Design-Based, Model-based and Model-Assisted Inference:

A Tutorial

By

Bill Warren

30 July 2004

A discussion paper produced for the Vegetation Resources Inventory Section
Resource Information Branch
Ministry of Sustainable Resource Management

1. Design-Based Estimation

1.1: Simple Random Sampling

The first step in design-based inference is the selection of the frame. In the words of Cochran (1963),

> "*Before selecting the sample, the population must be divided into parts which are called* ***sampling units****, or* ***units****. These units must cover the whole of the population and the must not overlap, in the sense that every element in the population belongs to one and only one unit.*"

In forestry, the sampling unit could be a tree and the population all the trees in a particular forest area. More usually, the sampling unit will be an area, with the forest area divided into squares, rectangles, pentagons, hexagons or even differently shaped and or sized polygons. Note that it is not possible to tessellate a planar area with circles; sampling with circle plots requires special treatment and is outside the scope of this document.

Let us suppose that the population consists on N such units and, of these, n are chosen for the sample. The simplest design-based approach is simple random sampling (*srs*). When the sample is chosen without replacement, there are $N!/n!(N-n)!$ ways of selecting the sample and, under simple random sampling, each of these has exactly the same chance of arising. Let the quantity of interest on the i^{th} unit be denoted by Y_i. In the design-based approach, there is no notion that these are random variables; they are no more than the observations associated with the sampling units. The sampling distribution is generated by the randomization process and is formed by the $N!/n!(N-n)!$ possible outcomes.

The population total is

$$Y=\sum^{N}Y_i \qquad [1]$$

and the population mean

$$\bar{Y} = \frac{\sum^{N} Y_i}{N}. \qquad [2]$$

The sample mean,

$$m=\sum^{n}Y_i/n \qquad [3]$$

is an unbiased estimator of the population mean, $\bar{Y}$, in the sense that its expectation, that is its average over the $N!/n!(N-n)!$ possible outcomes is $\bar{Y}$.

Let us define the population variance as

$$S^2 = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}{N-1}. \qquad [4]$$

The variance of the sampling distribution of the sample means, that is the average of the $(m-\bar{Y})^2$ over the $N!/n!(N-n)!$ outcomes, is $(1-(n/N))S^2/n$. Further, let $s^2=\sum^n(Y_i-m)^2/(n-1)$; then, in the same sense as above, s^2 is an unbiased estimator of S^2 and $(1-(n/N))s^2/n$ is an unbiased estimator of the sampling variance of the sample mean. (This is illustrated in Example 1 in the Appendix.) Often, n/N, will be close to 0.

Although the Y_i are not viewed as random variables, if the sample size n is sufficiently large, the sampling distribution will commonly be adequately represented by the normal, thus permitting the construction of reasonably accurate confidence limits. Exceptions can and do, arise.

1.2: Sampling With Unequal Probability

There are situations where sampling with unequal, but known, probability is expedient. Let p_i be the probability that the i^{th} unit is selected; $\sum^N p_i=1$. To avoid complications, sampling is carried out with replacement. There is, therefore, the possibility that the same unit will be selected more than once. Indeed, the probability that the i^{th} unit will be selected n times, i.e., comprise that whole sample, is $p^n_i>0$. There are then, with order of selection taken into account, N^n possible outcomes. The probability of obtaining Y_i, Y_j ... Y_r, say, in that order is $p_ip_j...p_r$.

$$\hat{Y} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i}{p_i}\right) \qquad [5]$$

is then an unbiased estimator of the population total Y, in the sense that the sum of all possible estimates, multiplied by the probability of their arising, is Y. The variance of the $\hat{Y}$ is

$$V(\hat{Y}) = \frac{1}{n}\sum_{i=1}^{n} p_i\left(\frac{Y_i}{p_i} - Y\right)^2 \qquad [6]$$

and an unbiased estimator of this variance is

$$V(\hat{Y}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i}{p_i} - \hat{Y}\right)^2 /(n-1). \qquad [7]$$

An illustration is provided in the Appendix, Example 2.

A special case arises if sampling is with probability proportional to size, that is in our case, unit area, a_i; $\sum^N a_i = A$, the area of the forest being surveyed. The $p_i = a_i/A$ and Y becomes

$$Y = \frac{A}{n} \cdot \sum \left(\frac{Y_i}{a_i} \right) \qquad [8]$$

with estimated variance

$$V(Y) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{AY_i}{a_i} - \hat{Y} \right)^2 / (n-1). \qquad [9]$$

Let $y_i = Y_i/a_i$, i.e., the volume per hectare on the i^{th} unit. Then

$$\hat{Y} = \frac{A \sum_{i=1}^{n} y_i}{n} = A\bar{y}, \qquad [10]$$

with estimated variance

$$S^2 = \frac{A^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}{n(n-1)}. \qquad [11]$$

The estimate of the mean volume per hectare is then

$$\frac{\hat{Y}}{A} = \bar{y}, \qquad [12]$$

with estimated variance

$$V(\bar{y}) = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n(n-1)}. \qquad [13]$$

These look like estimates from simple random sampling but it must be remembered that the units are being selected with probability proportional to size and the quantity being observed (or in some way determined) is the total volume for a unit.

The estimator given above is known as the Hansen-Hurwitz estimator. There is also the Horvitz-Thompson estimator, $\sum^n Y_i/\pi_i$, where π_i is the probability that the i^{th} unit is included in the entire sample as opposed to its selection at each draw, p_i. Then $\pi_i = 1-(1-p_i)^n \approx np_i$. This estimator will not be discussed further here.

1.3: Systematic Sampling

For simplicity, it will be supposed that N is a multiple on n, i.e., $N=kn$. In systematic sampling the N units are ordered in some manner and the sample consists of every k^{th} unit. If a random start is made from one of the first k units, there are k possible outcomes and the expectation of the sample mean, i.e., its average over all k possible outcomes, is the population mean, i.e., the estimator is design unbiased.

The estimator has a variance, namely the variance of the k possible estimates but, unless we make additional assumptions or have estimates from more than one random start, it is not possible to estimate this variance. This is essentially because, the selection of the first unit determines the selection of the remaining $n-1$ and the variability amongst these units does not necessarily reflect the variability of the sampling distribution. In effect, we have a sample of size 1 from a population of size k.

The estimate from a systematic sample is often, but not always, more precise (i.e., has smaller variance) than that from a simple random sample; the problem is that, without additional information or assumptions, this cannot be established for any particular case.

A plethora of methods have been suggested for dealing with systematic sampling, most of which are viable under some circumstances but not others. One approach will be discussed under Model-Based Estimation below.

1.4: Ratio Estimation under Simple Random Sampling

On each unit there are two measurable quantities, the Y_i observed on the sample of n units and X_i generally available on all N units. The ratio R is by definition, $\sum^N Y_i/\sum^N X_i$. Under simple random sampling the obvious estimator of R is

$$\hat{R} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i}, \qquad [14]$$

with the population total of Y_i estimated by $\hat{R}X$, where as before, $X = \sum^N X_i$.

In general, $\hat{R}$ is a biased estimator of R in that the average of the $N!/n!(N-n)!$ estimates does not equal R, although in most cases the bias will be small and, for practical purposes, negligible. Cochran notes that

*"... the ratio estimate is unbiased if the regression of **y** on **x** is a straight line through the origin. This means that **E(y/x)=βx**. In a finite population, the relation means that (a) if several units have exactly the same value of **x**, the mean of their **y**-values is **βx**, and (b) if a specific value of x occurs on only one unit in the population, the value of **y** for that unit is **βx**. These relations are unlikely to be satisfied exactly in a finite population".*

Cochran also observes "We do not posses exact formulas for the bias and the sampling variance of the estimate but only approximations that are valid in large samples. The approximation to the sampling variance of $\hat{R}$ is given as:

$$V(R) = \left(1 - \frac{n}{N}\right)\frac{1}{\bar{X}}\frac{\sum_{i=1}^{N}(Y_i - RX_i)^2}{n(N-1)} \qquad [15]$$

The sample-based estimator is

$$V(\hat{R}) = \left(1 - \frac{n}{N}\right)\frac{1}{\bar{X}}\frac{\sum(Y_i - \hat{R}X_i)^2}{n(n-1)} \qquad [16]$$

The average of the result of Equation 16 over all *N!/n!(N-n)!* outcomes will not, in general, equal the result of Equation 15 and neither equal the actual sampling variance obtained as the average of the R^2, less the square of the average of the

$$\hat{R} = E(\hat{R}^2) - E^2(\hat{R}).$$

These are illustrated in Appendix, Example 3.

1.5: Ratio Estimation under Sampling with Unequal Probability

Again R is defined as $\sum^N Y_i / \sum^N X_i = Y/X$. An obvious estimator of R is

$$\hat{R} = \frac{(1/n)\sum_{i=1}^{n}(Y_i / p_i)}{(1/n)\sum_{i=1}^{n}(X_i / p_i)} = \frac{\sum_{i=1}^{n}(Y_i / p_i)}{\sum_{i=1}^{n}(X_i / p_i)} \qquad [17]$$

Although the numerator and denominator are unbiased estimators of the population totals; Y and X, respectively, $\hat{R}$ is not, in general, an unbiased estimator of R.

As an estimator of the variance of $\hat{R}$, we may take the general form

$$v(\hat{R}) = \frac{1}{X^2} \frac{\sum \left(\frac{e_i}{p_i} - \left(\frac{1}{n} \sum \frac{e_i}{p_i} \right) \right)^2}{n(n-1)}. \qquad [18]$$

$$where: \quad e_i = Y_i - \hat{R}X_i$$

Since here

$$\hat{R} = \frac{\sum_{i=1}^{n} (Y_i / p_i)}{\sum_{i=1}^{n} (X_i / p_i)}, \qquad [19]$$

it follows that $\sum^n e_i/p_i=0$ and the variance estimate reduces to $(1/X^2)[(\sum^n (e_i/p_i)^2/n(n-1)]$. See appendix, Example 4 for an illustration.

The ratio estimate of the total is then $\hat{R}X$ with estimated variance $v(\hat{R})X^2$ and the mean per unit is $\hat{R}X/N$ with estimated variance $v(\hat{R})X^2/N^2$. Note, these variances are approximations and, in general biased.

2. Model-based Inference

2.1 Kriging the mean

In forestry we are commonly dealing with observations that are associated with locations and, to emphasize this, instead of Y_i we can write $Y_{(s_i)}$ to denote the observation at location s_i. Many of the qualities in which we are interested are spatially correlated, that is, observations at nearby locations will tend to be more similar than observations at more distant locations.

In design-based inference, the $Y_{(s_i)}$ are regarded as fixed and the locations are randomized. To emphasize the point Brus and de Gruijter (1993) write: $Y_{(S_i)}$. In contrast, in model-based inference, the sample locations are regarded as fixed and the observations are assumed to be a realization of some underlying stochastic process. Brus and de Gruijter then write: $\hat{Y}_{(S_i)}$. The distinction is important; I have elsewhere described it as follows (Warren 1998):

> *"In the design-based approach, expectations are taken over all possible sets of $\mathbf{s}_i$ of size $\mathbf{n}$; in the model-based approach, expectations are with respect to some assumed underlying stochastic process (or distribution). In both approaches, we imagine that our data are a subset of a realization of some stochastic process. In the design-based approach, we focus on this realization and look at what would be the average over all possible samplings of this realization;*

in the model-based approach, the set of locations is taken as fixed and we consider what would be the average of the (conceptual) realizations over this set".

The simplest assumption that can be made about the underlying stochastic process is that it is second order stationary, i.e., the expectation of $Y_{(Si)}$ is constant, μ, for all s_i and $Var(Y_{(Si)}-Y_{(Sj)})=Var(Y_{(Si)})+Var(Y_{(Si)})-2Cov(Y_{(Si)},Y_{(Sj)})$ depends only on the distance between the locations s_i and s_j.Thus $Var(Y_{(Si)}-Y_{(Sj)}) = 2\sigma^2(1-\rho))=2\gamma(h)$, say, where h is the distance between the locations. The function $\gamma(h)$ is called the semivariogram. Not all functions are valid as semivariograms. Possibly the most common and serviceable is the spherical model, namely $(\gamma(h)=0, \text{ if } h=0, \gamma(h)=c+b(3h/2a-1/2)(h/a)^3)$ if $0<h\leq a$, and $\gamma(h)=c+b$ if $h\geq a$. Thus $\gamma(h)$ increases, and the correlation $\rho(h)$ decreases, as h increases until $h=a$, known as range, after which $\gamma(h)$ remains constant, equal to the variance σ^2, and the correlation is zero.

This model is often used for interpolation, i.e., estimating the value of $Y_{(So)}$ at location *'so'* distinct from the sample locations. The estimator takes the form $\sum^n w_i Y_{(Si)} = \hat{Y}_{(So)}$ where the weights, w_i are chosen so that $E(\hat{Y}_{(So)})=Y_{(So)}$ and the variance of $v(\hat{Y}_{(So)}-Y_{(So)})$ is minimized. This leads to a system of $n+1$ linear equations in $n+1$ unknowns. The method can also be used to estimate the process mean, i.e., kriging the mean. Again the estimator takes the form $\sum^n w_i Y_{(Si)}=\hat{\mu}$, where the weights are chosen so that $E(\hat{\mu}) = \mu$ and $E(\hat{\mu}-\mu)$ is minimized. This again leads to a system of $n+1$, linear equations in $n+1$, unknowns.

It is emphasized that these expectations are evaluated with respect to the assumed stochastic process. It should also be noted that, since the data are regarded as a realisation of the process, the mean of the $Y_{(Si)}$ over N locations (the population mean, Y) will not necessarily equal the process mean, μ, and it is this latter that is being estimated. Also, the design-based S^2 differs from σ^2.

There is no requirement that the stochastic process be Gaussian (normal) however, if the process is Gaussian, linear estimation will be optimal (minimal variance).

The sample locations can be arbitrary. Having the sample locations on a regular grid, however, generally facilitates estimation of the variogram. Thus, there is no barrier to estimating the process mean, and its estimation variance, from a systematic sample.

In practice, the kriged mean should differ little from the ordinary sample mean. The estimation variance is, however, dependent on the degree of spatial correlation and, all other things being equal, the higher the correlation, i.e., if $\gamma_2(h)<\gamma_1(h)$, the greater the estimation variance. This is because, the greater the correlation, the smaller the amount of independent information. In contrast, the variance, calculated as if the systematic sample were a simple random sample, is, in expectation, unaffected by the level of correlation.

If there is no spatial correlation, or if the distance between all sample locations exceed the range of the variogram, the kriged mean and its estimation variance will be numerically

the same as the sample mean and its variance under an assumption of simple random sampling. Although numerically equivalent, the estimates differ conceptually. The estimation variance refers to the variation that would arise from different realizations of the same process, where as the *srs* variance refers to the variation that would arise from different randomizations of the same realization. With no spatial correlation, it would be difficult, if at all possible, to distinguish between these two situations. This would not in general, be the case in the presence of spatial correlation. Thus, although under non-random and, in particular, systematic sampling, one may obtain a model-based estimate of a mean and its variance, these cannot be given the same interpretation as the design-based estimates.

2.2: Model-Based Ratio Estimation

Let $Y_i=RX_i+e_i$. The e_i are assumed to be independent random variables with mean 0 and variance σ^2_i. Since the variability of the Y_i commonly increases with its magnitude, it can then be assumed that $\sigma^2_i=kX^d_i$ *(where d is a maximum likelihood coefficient that defines the nature of the relationship between variance in Y_i in relation to X_i)*. In a forestry context, we may think of the Y_i as the actual net volume of a tree and X_i as a visual estimate of that volume.

$$\text{Let } \hat{R}=\frac{\sum_{i=1}^{n} w_i Y_i}{\sum_{i=1}^{n} w_i X_i}=R+\frac{\sum_{i=1}^{n} w_i e_i}{\sum_{i=1}^{n} w_i X_i}. \qquad [20]$$

It follows that, whatever the system of weights, w_i, $E(\hat{R})=R$, that is, $\hat{R}$ is a model unbiased estimator of R. The variance of $\hat{R}$ is then:

$$V(\hat{R})=\frac{\sum_{i=1}^{n} w_i^2 kX_i^d}{\left(\sum_{i=1}^{n} w_i X_i\right)^2} \qquad [21]$$

Given d, $V(\hat{R})$ is minimized with w_i proportional to X_i^{1-d}. Then

$$\hat{R}=\frac{\sum_{i=1}^{n} X_i^{1-d} Y_i}{\sum_{i=1}^{n} X_i^{2-d}} \qquad [22]$$

An obvious model-unbiased estimator of $V(\hat{R})$ is:

$$V(\hat{R}) = \frac{\sum w_i^2 e_i^2}{\sum w_i X_i} = \frac{\sum X_i^{2-2d} e_i^2}{\left(\sum X_i^{2-d}\right)^2} \qquad [23]$$

$$where: \quad e_i = Y_i - \hat{R}X_i.$$

Special cases occur with $d=0$, i.e., $V(Y_i)$ constant, $d=1$, i.e., $(V(Y_i)$ is proportional to the mean, and $d=2$, i.e., the standard deviation of the Y_i is proportional to the mean. Then for:

d=0
$$\hat{R} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} \qquad [24]$$

d=1
$$\hat{R} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i} \qquad [25]$$

d=2
$$\hat{R} = \frac{\sum (Y_i / X_i)}{n} \qquad [26]$$

In the above, d has been assumed to be known. If we add the assumption that the e_i are normally distributed, the method of maximum likelihood can be used to estimate d; indeed all three parameters, R, k, and d may be estimated, R from d and k from R and d. Unfortunately, there is no closed form expression for d and it has to be estimated iteratively. For practical purposes, it may suffice to guess a value for R, assume a value for d, calculate k and evaluate the likelihood. By repeating this for values of d, one can construct a likelihood profile and visually determine the value of d for which the profile is maximized. It may well turn out that the profile has a relatively broad plateau and, in effect, any value of d corresponding to the plateau would be consistent with the data.

The above can be placed in the context of the general linear model namely:

$$Y_i = a + bX_i + e_i \qquad [27]$$

Where, the e_i are taken to be independent random variables with expectation 0 and variance v_i. (The assumption of independence can be removed if necessary.) In the case of ratio estimation, $a=0$ and we have written b as R.

If v_i is constant, the best linear unbiased estimator of R is $\sum^n Y_i X_i / \sum^n X_i^2$. If we allow for unequal probability of selection, the estimator becomes $\sum^n (Y_i X_i / p_i) / \sum^n (X_i^2 / p_i)$. If the v_i are not constant, the general linear model estimator of R (under simple random sampling) is:

$$\hat{R} = \frac{\sum_{i=1}^{n} (Y_i X_i / v_i)}{\sum_{i=1}^{n} (X_i^2 / v_i)} \qquad [28]$$

This is equivalent to regressing $Y_i^* = Y_i/\sqrt{v_i}$ on $X_i^* = X_i/\sqrt{v_i}$. Thus, with unequal probability of selection, the estimator would be:

$$\hat{R} = \frac{\sum_{i=1}^{n} (Y_i^* X_i^* / p_i)}{\sum_{i=1}^{n} (X_i^{*2} / p_i)} \qquad [29]$$

$$i.e. \quad \hat{R} = \frac{\sum_{i=1}^{n} (Y_i X_i / p_i v_i)}{\sum_{i=1}^{n} (X_i^2 / p_i v_i)}$$

This is illustrated numerically in the Appendix (Example 5).

Compare this with the design-based estimator under unequal probability given above, namely:

$$\hat{R}_d = \frac{\sum_{i=1}^{n} (Y_i / p_i)}{\sum_{i=1}^{n} (X_i / p_i)} \qquad [30]$$

Thus the design-based estimator would be equivalent to the model-based estimator with v_i proportional to X_i. This leads us to the subject of model-assisted inference.

3. Model-Assisted Inference.

Conquest (2003) writes:

> *"In a model-assisted approach] ancillary data is made use of by specifying what is called a* ***super-population model*** *between the ancillary variables and the response(s) of interest. The model is then used to either suggest a design or an estimator. In the estimation case, although the model suggests the estimators, the estimators are evaluated with design-based priorities (Opsomer et al. 2001). For example, regression estimation uses a covariate within the context of a linear model to obtain a more precise estimate of a total or a mean of the population. The regression estimator, although slightly biased, results in a much higher (design-based) precision than the estimator that does not take advantage of the ancillary data. In the design case, the model is used to suggest a probability sampling-design that takes advantage of the relationship between the ancillary variables and the response. If the model holds, these strategies can help the sampler estimate parameters or construct optimal sampling designs".*

The conventional design-based estimator is $\sum^n(Y_i/p_i)/\sum^n(X_i/p_i)$. As shown above, this makes the tacit assumption that the variance of the e_i (and thus the variance of the Y_i) is proportional to X_i. This is the optimal situation for regression estimation. But what if v_i is constant or proportional to X_i^2, or, perhaps totally unrelated to X_i? The appropriate estimator is, as given below:

$$R = \frac{\sum_{i=1}^{n}(Y_i X_i / p_i v_i)}{\sum_{i=1}^{n}(X_i^2 / p_i v_i)} \qquad [31]$$

Since, under model-assisted inference, the properties are evaluated with respect to the design, the suggested estimator for $V(Y)$ is:

$$V(Y) = \frac{\sum_{i=1}^{n}\left[\frac{e_i}{p_i} - \left(\frac{1}{n}\sum_{i=1}^{n}\frac{e_i}{p_i}\right)\right]^2}{n(n-1)} \qquad [32]$$

And, since $Y=RX$, the estimator for $V(\hat{R})$ would be:

$$V(\hat{R}) = \frac{1}{X^2}\cdot\frac{\sum_{i=1}^{n}\left[\frac{e_i}{p_i} - \left(\frac{1}{n}\sum_{i=1}^{n}\frac{e_i}{p_i}\right)\right]^2}{n(n-1)} \qquad [33]$$

where $e_i = Y_i - \hat{R}X_i$.

Note that $\hat{R}$ is evaluated from the Y_i^* and X_i^* whereas the e_i use the Y_i and X_i. Note also that, in the special case that v_i is proportional to Xi, $\sum^n e_i/p_i=0$, but this does not hold in general.

While both estimators are model-unbiased, neither is design-unbiased, although the bias should be small. The mean-square error would then normally be the basis of comparison. The above expression for the variance is also an approximation. A numerical example is given in the Appendix (Example 6).

4. Observations

There is a potential problem in applying model-based inference when sampling with replacement. The assumption is that the e_i are independent, but there is a chance that the same unit will occur more than once in a sample. This would not be a problem when, for example, the unit is an area (polygon) and more than one location (plot) is sampled within the polygon. We then have two (or more) independent estimates for the unit. Fortunately, the case where the unit is, say, a tree and only one observation for Y_i is possible, will, in practical situations, occur rarely. In the examples of the Appendix, with N only 6 and $n=2$, there is almost a 20% chance that this will happen, but it is only when such does happen that a problem arises and the methodology has to be extended to deal with it. This is beyond the scope of this document.

The theory of regression estimation parallels that of ratio estimation. When the auxiliary variable, X_i, is available for all sampling units, there is, of course, no sampling error for the total, or mean, of the X_i. The precision of the estimate of the total, Y, is then primarily governed by one's ability to estimate the ratio, R, or the regression parameters. The sampling design should then be focused on estimation of the relationship.

For example, there would appear to be very little difference between the results obtained from systematic sampling from an ordered list and sampling with probability proportional to size with replacement but, while these will likely be better than simple random sampling, better strategies might exist and be derived from an appropriate model. For example, in the model-based approach given above, the minimum variance of R, given d is proportional to: $1/(\sum X_i^{2-d})^2$. Thus if d were less than 2, the strategy of purposively sampling the units with the larger X_i would be advantageous. However, if d were greater than 2, i.e., the variance going up by more than the square of the observation, admittedly a relatively rare circumstance, the better strategy would be to purposively sample the smaller X_i. Note that purposive sampling is not the same as subjective sampling.

References

Brus, D.J. and de Gruijter, J.J. 1993. Design-based versus model-based estimates of spatial mean; theory and application in environmental soil science. Environmetrics 4:123-152.

Cochran, W.G. 1963. Sampling Techniques 2nd Ed. Wiley, New York, NY

Conquest, L.L. 2003. Model-assisted sampling approaches in the sampling of natural resources. American Statistical Association, Section of Statistics and the Environment, News letter: V.5, No 1.

Opsomer, F.J., Moisen. G.G. and Kim, J.Y. 2001. Model-assisted estimation of forest resources with generalised additive models. Proceedings of the Section on Survey Research Methods, American Statistical Association, Art. #00369.

Warren, W.G. 1998. Spatial analysis of marine populations: factors to be considered. Canadian Special Publication, Fish and Aquatic Science. 125:21-28.

APPDENDIX

The following examples are not intended to be realistic; their role is to illustrate and hopefully, clarify the methodology. No generalisation of the relative performance of the estimators should, therefore, be made.

Example 1. (Simple random sampling)

Let $N=6$, $n=2$, $Y_i=[61.6\ 56.3\ 50.8\ 53.5\ 43.0\ 67.0]$
The population total $Y = 332.2$, mean $(Y_i) = 332.1/6 = 55.36$
And $S^2 = (61.62 + 56.32 + ... + 67.02 - 332.22/6)/5 = 70.46$

There are $6!/2!4! = 15$ possible samples, thus

Combination	sample values	Y_i+Y_j	$\hat{Y}$	$(Y_i-Y_j)^2/2$
1,2	61.6 56.3	117.9	353.7	14.045
1,3	61.6 50.8	112.4	337.2	58.320
1,4	61.6 53.5	115.1	345.3	32.805
1,5	61.6 43.0	104.6	313.8	172.980
1,6	61.6 67.0	128.6	385.8	14.580
2,3	56.3 50.8	107.1	321.3	15.125
2,4	56.3 53.5	109.8	329.4	3.920
2,5	56.3 43.0	99.3	297.9	88.445
2,6	56.3 67.0	123.3	369.9	57.245
3,4	50.8 53.5	104.3	312.9	3.645
3,5	50.8 43.0	93.8	281.4	30.420
3,6	50.8 67.0	117.8	353.4	131.220
4,5	53.5 43.0	96.5	289.5	55.125
4,6	54.5 67.0	120.5	361.5	91.125
5,6	43.0 67.0	110.0	330.0	288.00
Total		1661.0		1055.00

Y would be estimated as $6(Y_i+Y_j)/2 = 3(Y_i+Y_j)$. Now, $3x1661.0/15 = 332.2$, i.e., $E(6m) = Y$.

With $n=2$, $\sum^n(Y_i-m)^2/(n-1) = (Y_i-Y_j)^2/2$. Thus the average of the $s^2 = 1055.0/15 = 70.46$ - illustrating the un-biasedness of s^2 as an estimator of S^2.

Finally, the variance of the estimates of m is:

$[(117.9^2+112.4^2+...+110.0^2)/4]/15-55.36^2 = 23.48$ which may be compared with: $(1-n/N)S^2/n = (2/3)70.46/2 = 23.48$, illustrating that $(1-n/N)s^2/n$ is an unbiased estimator of the variance of the sample mean.

Example 2. Sampling with unequal probability

We use the same observations, Y_i, as Example 1 but with selection probabilities p_i= [.20, .24, .16, .10, .05, .25]; thus:

$$Y_i/p_i = [308.0\ 234.583\ 317.5\ 535.0\ 860.0\ 268.0].$$

With $N=6$ and $n=2$ there are 21 distinct outcomes, however, it will be convenient to list the $Y_i/p_i+Y_j/p_j$ with respect to order, in a 6 by 6 array, thus

order	1	2	3	4	5	6
1	616.000	542.583	625.500	843.000	1168.000	576.000
2	542.583	469.166	552.083	769.583	1094.583	502.582
3	625.500	552.083	635.000	852.500	1177.500	585.500
4	843.000	769.583	852.500	1070.00	1395.000	803.000
5	1168.000	1094.583	1177.500	1395.00	1720.000	1128.000
6	576.000	502.583	585.500	803.000	1128.000	536.000

The estimates of Y are $1/n$, i.e., ½ of these values. The corresponding probabilities of occurrence, p_ip_j, are:

order	1	2	3	4	5	6
1	.0400	.0480	.0320	.0200	.0100	.0500
2	.0480	.0576	.0384	.0240	.0120	.0600
3	.0320	.0384	.0256	.0160	.0080	.0400
4	.0200	.0240	.0160	.0100	.0050	.0250
5	.0100	.0120	.0080	.0050	.0025	.0125
6	.0500	.0600	.0400	.0250	.0125	.0625

The expected value of Y is then the product of $(1/2)(Yi/pi+Yj/pj)$ and the corresponding p_ip_j, i.e., (1/2)(0.0400x616.0+0.0480x542.583+…+0.0625x536.0) = 332.2, thus illustrating the unbiasedness of the estimator.

The variance of the $\hat{Y}$ is then $E(\hat{Y}^2)-E^2(\hat{Y})$. Thus

$$E(\hat{Y}) = (1/4)(0.0400x616.0^2+0.0480x542.583^2+\ldots+0.0625x536.0^2) - 332.2^2$$
$$=10755.$$

The variance estimates are $\sum^n(Y_i/p_i-\hat{Y})^2/n(n-1)$ which, with $n=2$, becomes $(Y_i/p_i-Y_j/p_j)^2/2$.

The estimates are then:

order	1	2	3	4	5	6
1	0.0	2695.004	45.125	25764.500	152352.000	800.000
2	2695.004	0.0	8437.587	45125.088	195573.010	558.369
3	45.125	8437.587	0.0	23653.125	147153.130	1225.125
4	25764.50	45125.088	23653.125	0.0	52812.500	35644.500
5	152352.000	195573.010	147153.130	52812.500	0.0	175232.000
6	800.000	558.369	1225.125	35644.500	175232.00	0.0

The expected value of the variance estimates is then:

0.0+2695.005x0.0480+…+175232.0x0.0125+0 = 10755.0

thus illustrating the unbiasedness of the variance estimator.

Example 3. Ratio Estimation with Simple Random Sampling

We now introduce an auxiliary variable X_i = [33.7 25.1 24.2 27.6 19.9 35.6]. Thus

$$R = \sum^N Y_i / \sum^N X_i = 332.1/166.1 = 2.0$$

The possible estimates of R ($\sum^n Y_i / \sum^n X_i$) are then

Combination	Ratio computation	Ratio
(1,2)	117.9/58.8 =	2.0051
1,3	112.4/57.9 =	1.9413
1,4	115.1/61.3 =	1.8777
1,5	104.6/53.6 =	1.9515
1,6	128.6/69.3 =	1.8557
2,3	107.1/49.3 =	2.1724
2,4	109.8/52.7 =	2.0835
2,5	99.3/45.0 =	2.2067
2,6	123.3/60.7 =	2.0313
3,4	104.3/51.8 =	2.0135
3,5	93.8/44.1 =	2.1270
3,6	117.8/59.8 =	1.9699
4,5	96.5/46.5 =	2.0316
4,6	120.5/63.2 =	1.9066
5,6	110.0/55.5 =	1.9820

The expected value of $\hat{R}$ is then (2.0051+1.9413+…+1.9820)/15 = 2.0104. Thus, $\hat{R}$ is a biased estimator of R but, as expected, the bias is small (0.0104).

The variance of the sampling distribution of $\hat{R}$ is

$$(2.0051^2+1.9413^2+\ldots 1.9820^2)/15 - 2.0104^2 = 0.009789.$$

An approximation to this variance is given by $(1-n/N)\sum N(Y_i-RX_i)^2/n(N-1)\,\overline{X}$. Here the residuals e_i are Y_i-RX_i=[5.8 -6.1 -2.4 1.7 -3.2 4.2] and $\overline{X} = 166.1/6 = 27.683$, whence:

$$(1-2/6)107.38/(2x5x27.6832) = 0.009341.$$

The 15 sample-based estimates of $V(\hat{R})$ are $(1-n/N)\sum^n (Y_i - \hat{R}X_i)/n(n-1)\,\overline{X}^2$, i.e., [0.0310 0.0127 0.0024 0.05151 0.0008 0.0027 0.0139 0.0007 0.0246 0.0037 0.0004 0.0085 0.0058 0.0007 0.0110] with expected value (0.0310+0.0127+…+0.0110)/15 = 0.00893.

Thus the actual sampling variance of the $\hat{R}$, the expected value of its sample estimator and the approximation differ somewhat, being 0.009789, 0.00893 and 0.00934, respectively.

Example 4. Ratio Estimation with Unequal Probability

We continue with the same Y_i, X_i and p_i of the previous examples. The estimates of R, i.e., the

order	1.	2.	3.	4.	5.	6.
1.	1.8279	1.9869	1.9562	1.8965	2.0618	1.8527
2.	1.9867	2.2430	2.1580	2.0221	2.1779	2.0349
3.	1.9562	2.1580	2.0992	1.9953	2.1438	1.9939
4.	1.8965	2.0221	1.9953	1.9384	2.0697	1.9192
5.	2.0618	2.1779	2.1438	2.0697	2.1608	2.0873
6.	1.8527	2.0349	1.9939	1.9192	2.0873	1.8820

The expected value of $\hat{R}$ is then

$$(1.8279\text{x}0.0400+1.9869\text{x}0.0480+\ldots+1.8820\text{x}0.0625) = 2.0025$$

thereby illustrating that $\hat{R}$ is a (slightly) biased estimator of R.

The variance of $\hat{R}$ is

$$(1.8279^2\text{x}0.0400+1.9869^2\text{x}0.0480+\ldots+1,8820^2\text{x}0.0625) - 2.0045^2 = 0.012365$$

The sample estimates, $\sum^n (Y_i/p_i - \hat{R}\, X_i/p_i)^2/n(n-1)X^2$ are

order	1.	2.	3.	4.	5.	6.
1.	0.0	0.0260	0.0169	0.0048	0.0563	0.0006
2.	0.0260	0.0	0.0029	0.0193	0.0017	0.0172
3.	0.0169	0.0029	0.0	0.0089	0.0017	0.0092
4.	0.0048	0.0193	0.0089	0.0	0.0476	0.0010
5.	0.0563	0.0017	0.0017	0.0476	0.0	0.0010
6.	0.0006	0.0172	0.0092	0.0010	0.0310	0.0

The expected value of the variance estimates is then:

$(0.0\text{x}0.0400+0.0260\text{x}0.0480+\ldots+0.0310\text{x}0.0125+0.0\text{x}0.0625) = 0.0106$, slightly less than the actual variance, 0.0124.

Example 5. Model-Based Ratio Estimator

In this example simple random sampling will be assumed. The variance σ_i^2 is assumed to be proportional to X_i^d. With $N=6$ the likelihood profile is almost flat. There is a maximum for $d \approx 1.3$. For convenience, d will be taken as $3/2$,then

$$\hat{R} = \sum^n X_i^{3/2} Y_i / \sum^n X_i^{5/2}.$$

With the same data as before, we obtain:

Combination	Model-based ratios
1,2	2.0202
1.3	1.9523
1,4	1.8804
1,5	1.9726
1,6	1.8553
2,3	2.1718
2,4	2.0871
2,5	2.2043
2,6	2.0468
3,4	2.0160
3,5	2.1285
3,6	1.9802
4,5	2.0405
4,6	1.9804
5,6	2.0013

The expected value is, then, $(2.0202+1.9523+...+2.0013)/15 = 2.0177$.

The variance of the estimator is:

$$(2.0202^2+1.9523^2+...+2.0013^2)/15 - 2.0177^2 = 0.0095$$

Taking $e_i = Y_i - \hat{R} X_i$ and substituting in $\sum^n X_i^{2-2d} e_i^2 / \sum^n (X_i^{2-d})^2$, we obtain variance estimates 0.0213, 0.0091, 0.0015, 0,0135, 0.0004, 0.0026, 0.0116, 0.0008, 0.0160, 0.0032, 0.005, 0.0058, 0.0061, 0.0004, 0.0093, with average 0.0068. The estimates thus tend to be lower than the actual variance, 0.0095, suggesting that an alternative variance estimator might be more appropriate.

Example 6. Model-Assisted Ratio Estimation with Unequal Probabilities.

Here the estimator is $\sum^n(X_iY_i/p_iv_i)/\sum^n(X_i^2/p_iv_i)$ and, as in Example 5, we take vi proportional to X_i^d with $d=3/2$. Recall that we are here sampling with replacement.

The estimates are:

order	1.	2.	3.	4.	5.	6.
1.	1.8279	2.0016	1.9674	1.8991	2.0791	1.8523
2.	2.0016	2.2430	2.1573	2.0250	2.1764	2.0505
3.	1.9674	2.1673	2.0992	1.9978	2.1450	2.0043
4.	1.8991	2.0250	1.9987	1.9384	2.0784	1.9208
5.	2.0791	2.1764	2.1450	2.0784	2.1608	2.1020
6.	1.8523	2.0505	2.0043	1.92308	2.1020	1.8820

The expected value is:

$$(1.8279x0.0400+2.0016x0.0480+...+1.8820x0.0625) = 2.0084$$

and the variance:

$$(1.8279^2x0.0400+2.0016^2x0.0480+...+1.8820^2x0.0625) - 2.0084^2 = 0.0125.$$

The variance estimates, obtained as $\sum^n(e_i/p_i-(1/n)(\sum^n e_i/p_i)^2/n(n-1)X^2$, with $e_i=Y_i-\hat{R}X_i$ are:

order	1.	2.	3.	4.	5.	6.
1.	0.0	0.0269	0.0171	0.0047	0.0508	0.0006
2.	0.0269	0.0	0.0029	0.0198	0.0016	0.0176
3.	0.0171	0.0029	0.0	0.0091	0.0016	0.0091
4.	0.0047	0.0198	0.0091	0.0	0.0462	0.0010
5.	0.0508	0.0016	0.016	0.0462	0.0	0.0271
6.	0.0006	0.0176	0.0091	0.0010	0.0271	0.0

The expectation for variance is:
(0+0.0269x0.0480+...+0.0271x 0.0125+0.0) = 0.0105.